A TELUGU MORPHOLOGICAL ANALYZER

Uma Maheshwar Rao G., Amba Kulkarni P. and Christopher Mala

Center for Applied Linguistics and Translation Studies
University of Hyderabad
Hyderabad, India
guraohyd@yahoo.com, ambapradeep@gmail.com, efthachris@gmail.com

Abstract

A Morphological Analyzer (MA) is a program which compiles and analyses words of a natural language into their roots and their constituent morpho-syntactic elements along with their attributes. The present paper demonstrates computational implementation of a Morphological Analyzer for Telugu. The algorithm used to build this MA is theoretically justified and is practically executed for Telugu in the context of Modern Standard Written variety. The present proposal is a demonstration of the optimal organization of linguistic database and its performance in computational environment by ensuring high precision and coverage in the parsing of word-forms. The current MA engine's coverage may range between 95-97% on a variety of corpora (3 million word length corpus).

Introduction:

It is a well known fact that the morphology of Telugu is not only rich in terms of the density of word-forms produced for a given root/stem but also diverse in the morphological strategies that are usually employed. The treatment of morphology from the computational point of view has become very important not because of the sheer number and variety of forms produced and the amount of information that is stored and available with the word-forms but because of its role as an important component of any NLP task.

A language like Telugu is regarded as morphologically rich wherein the words are formed not only from concatenation of one or more stems/roots plus one or more suffixes or morphological elements [Krishnamurti, 1985] but also involve multiple derivation. The complexity of morphology requires a more sophisticated morphological analyzer which is an essential tool in Machine Translation, Information Extraction, Information Retrieval, Speech synthesis and other NLP applications[Rao et.al. 2006].

Here, we propose a detailed description of the database and its compilation for the purpose of morphological analysis using Word and Paradigm Model[Hockett, 1954]. The present proposal is a demonstration of the optimal organization of linguistic database and an efficient use of computational space and performance ensuring precision and coverage in the parsing of word-forms. The present paper discusses the organization of the relevant linguistic data for building well-equipped, theoretically satisfiable and practically usable Morphological Analyzer for Telugu in the context of Modern Standard Written Variety.

The present Telugu MA is integrated into Telugu-Tamil and Telugu-Hindi Machine Translation systems. The Telugu MA displays the results of a word-form in one or more of 2800 morphological categories[Rao et.at. 1999, 2007]. The input and output specification follows the SSF format. It works both as stand-alone as well as system deployable module. It is tested on a variety of corpora with the coverage of 95-97%.

Methodological Foundation for Building MA:

In order to build a MA and a Morphological Generator (MG) for morphologically complex languages, it is essential that these tools must be built based on an appropriate theoretical foundation of morphology. The morphological model selected for this purpose must satisfy certain requirements such as it should be an evolutionary system allowing further modifications in the future without a drastic change in the already existing system. It should be a transparent system where an end-user can use it or modify it according to his needs. Over and above, it should be a modular system, wherein a change in a specific module would not affect the overall system. In other words, the modules of linguistic database, the meta-linguistic database and the computational modeling need to be neatly organized into distinct modules.

Morphological Models:

There are various models proposed to account for the morphological processes involved in human languages. Hockett [1954] distinguished the following morphological approaches which he called Item and Arrangement, Item and Process and Word and Paradigm. In the Item and Arrangement (IA) model, one analyzes word forms as sequences of concatenated morphemes. This involves the description of a specified set of morphemes (items) and their possible configurations in which the items can co-occur (arrangement). It subsumes no notion of allomorphs. It is Morpheme based Morphological analysis. In the Item and Process (IP) model one assumes that a word form is obtained by the result of applying rules which modify a root or stem in order to produce a new one. The morphology that is considered here involves a set of processes which act on roots/stems and generate word forms. It uses the notion of allomorphs. It is Lexeme based morphological analysis. In the Word and Paradigm (WP) model the inflexional paradigm is the central notion. A paradigm is considered as a set of morpho-syntactically related word forms each of these is associated with the given lexeme. Each word-form is a maximal projection. WP assumes a word as a formally realized word form, a functional projection of the root/stem (lexeme) according to the specifications of the formative elements in the paradigm. WP treats words as whole words and each of these is considered as a functional projection of its lexeme. The WP model is not strictly considered as word based since lexeme is still the base for the derivation of its corresponding word forms.

The above mentioned three models represent the combinatorial approach involving the segmentation of a string into a base and its constituent affixe(s). Ford and Singh(1985) propose a relational approach to morphology which focuses on the word and discards the notion of morpheme. This type of Morphology alternatively known as whole-word morphology is essentially a list of exhaustive morphological relations expressed in the form of morphological strategies.

The Word and Paradigm (WP) Model is considered as a better suited model for agglutinative language like Telugu for a computational implementation. Instead of stating rules to combine various morphemes into word forms, or to analyze word forms into a stem/root plus various morphemes, the WP model assumed here involves an exhaustive collection of each and all word forms relatable to a lexeme. This requires the identification of all the inflectional categories in the language, identification of conjugational and declensional classes of paradigms and finally listing of all paradigmatic members for each of these.

The term Paradigm refers to an exhaustive set of morpho-syntactically related word forms of a given lexeme. An inflectional category is a morpho-syntactic function expressed as a word form. A collection of such inflectional categories is often referred to as characteristic conjugations of verbs and declensions of nouns and so on. Various conjugational or declensional classes refer to the morphophonemic distinctions or similarities in the formation of word forms. Each word form in a paradigm is considered as a formal expression of the root/stem (lexeme) associated with a morpho-syntactic function.

The word forms of a typical lexeme are organized conveniently into tables, by classifying them according to the shared inflectional categories such as tense, aspect, modals and optionally gender, number, person in verbs and gender, number, person and case in nouns and so forth. There is an added advantage of using paradigms since it is easier to organize, modify and improve upon the data at a later stage.

1. Organization of Data:

The current morphological analyzer works on Word and Paradigm method and requires the following basic resources or the morphological data base that is described below. These resources are of three types:

1.1 Paradigm Table:

There are six morphological or lexical categories in Telugu viz. nouns, verbs, pronouns, number words, adjectives and locative nouns (traditionally included in the class of adverbs but re-named as separate category of nouns of space and time). Lexemes or root/stem of these categories systematically inflect for various morpho-syntactic functions. The so called inflectional categories associated with these are characteristic of these categories. These roots or stems which lack systematic characteristic inflexion simply do not involve any morphology and simply get listed in the Lexicon hence they come under indeclinables. The task of MA is to analyze a given word in terms of one of these six inflecting lexical categories or as an indeclinable. An exhaustive listing of word-forms in each category require to classify root/stem into various morphophonemic variation. They exhibit in word formation (formation of paradigmatic forms). The following table depicts the categories and number of morphophonemic classes and number of paradigmatic categories or morphological categories in each case as shown in the table-1.

S.No	Category	Morphophonemic Classes	Morphosyntactic categories	Paradigmatic data

1	Noun	52	200	10400
2	Pronoun	20	200	4000
3	Number word	10	22	220
4	Locative Noun	8	54	432
5	Adjective	13	36	468
6	Verb	34	930	31620
7	Indeclinable	6	7	42

Table-1: Paradigmatic Categories

1.1 A Lexicon :

A lexicon is a dictionary containing a list of roots/stems, each with its lexical category and paradigm type (Rao et.al 2007). Every proposed analysis of a word shall be validate by finding a match against a root word and the corresponding category and the paradigm type. It is organized in the form of a simple linear, non-hierarchical sequence of the root, delimiter, lexical category, delimiter and paradigm type as in Table 2. In addition to the paradigmatic database every lexeme that exists in the language shall be exhaustively listed in the morph lexicon. Currently there are about 75000 root words.

S.No	**Root/Stem form**	**Lexical Category (lcat)**	**Paradigm Type**
1	*“winu”*	“v”	*“koVnu”*
2	*“maMwri”*	“n”	*“gaxi”*
3	*“welika”*	“adj”	*“lewa”*
4	*“appudu”*	“adv”	*“appudu”*
5	*“iwadu”*	“pn”	*“vAdu”*

Table 2: Lexicon

1.2 Feature Value Table:

The Feature Value Table is essentially a list of affixes with their morpho-syntactic feature values like gender, number, person and the relevant morphological category information stored in the form of a table (Rao et.al 2007). Feature value Table contains lcat, affix and case associated with nouns and pronouns, the tense, aspect and modal categories with or without gender, number and person associated with verbs; the affixes associated with adjectives and locative nouns as shown in Table 3.

S.No	**Rule No.**	**lcat**	**Affix**	**Gender**	**Number**	**Person**
1	719	v	*iwi*	m	pl	2
2	653	n	*wopAtu*	null	sg	null

3	1649	pn	*lekuMdA*	null	pl	null
4	963	adj	*ti*	null	pl	null

Table 3: Feature Value Table

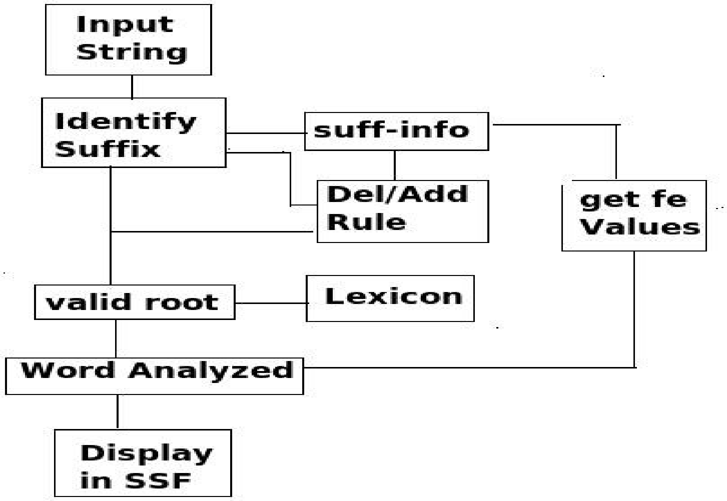
1.3 Morphophonemic Rules Set:

Word forms are analyzed by an exhaustive set of deletion/addition rules as in table-4 (Rao et.al 2007). The Morphop rule set is an exhaustive rule set, essentially a combination of concatenation processes which add the desired suffix to the given root/stem and which itself is appropriately modified by the relevant deletion rules. Both the add rule and deletion rule may apply vacuously in case the value of the character string to be added or deleted is null.

S.No	Delete Rule from wordform	Concatenation Add Rule to stem	Paradigm Type	Rule No.
1	*A*	*u*	*vAdu*	1649
2	*IsAdA*	*iyyi*	*wiyyi*	653
3	*akuMdA*	*u*	*poVg?du*	719
4	*NNilekuMdA*	*du*	*snehiwudu*	963

Table 4 : Morphophonemic Rules

2. Implementation of Telugu Morphological Analyzer:



A valid alpha numeric string bounded by spaces is taken as input for the Morphological Analyzer. The MA engine tries to identify the suffix by stripping each character from right to left and tries to match in the suff-info table, where deletion-addition rules are listed exhaustively. By matching the stripped string, the deletion and addition rules are applied on the word form. Each root that is formed by applying the del-add rule are validated by cross verifying from the lexicon. If the category and the paradigm are matched, retrieves the grammatical feature values from the Fe file. The base/root and its feature structure are displayed in SSF format.

3. Conculsion: The current MA analyses every valid Telugu word irrespective or its inflectional and derivational morphological formation. The derivational module is the most important and crucial component as it is groomed to analyze every potential word and not limiting itself with the attested words.

Transliteration Scheme using wx-notation:
a A i I u U q Q eV e E oV o O M H;
k K g G f c C j J F t T d D N w W x X n p P b B m y r rY l lY lYY v S R s h

References:

Hockett, C.F. 1954. *Two Models of Grammatical Description*. *Word*, 210-231. [= Readings in Linguistics Vol. I 386-399]

Krishnamurti. Bh and J.P.L. Gwynn. 1985. *A Grammar of Modern Telugu*. New Delhi. Oxford University Press.

Uma maheswara Rao, G. 1999. *A Morphological Analyzer for Telugu* (electronic form). Hyderabad: University of Hyderabad.

Uma maheswara Rao, G. and Amba Kulkarni, P. Christopher, Mala. 2007. *Morphological Analyzer and Its Functional Specifications for IL-ILMT System.* CALTS, Hyderabad: University of Hyderabad.

Uma maheswara Rao, G. and Amba Kulkarni, P. 2006. *Computer Applications in Indian Languages,* Hyderabad: The Centre for Distance Education, University of Hyderabad.